数字人文“一手证据”循证范式研究：基于《鲍氏国策》的共词分析

魏志鹏[1,2]，赵悦言[3]，杨克虎[1,2]，周文杰[1,3*]

（1. 兰州大学 循证社会科学交叉创新实验室，兰州 730030；2. 兰州大学 基础医学院循证医学中心，兰州 730030；

3. 西北师范大学 商学院，兰州 730070）

摘　要：［目的 / 意义］蕴藏于原始文献典籍的“一手证据”是展开数字人文研究的重要途径。［方法 / 过程］本文以《鲍氏国策》为例，基于自然语言处理技术，以共词分析方法为突破口，比较全面地展开了数字人文研究者如何基于来自原始文献的“一手证据”展开系统化的研究。［结果 / 结论］本研究针对《鲍氏国策》中“一手证据”的提取，从词频统计、停用词的去除、词义模糊词的识别与剔除等方面，展开了基于词语展开分析时若干基础指标的获取过程；进而，本研究以《鲍氏国策》为例，提供了数字人文研究中，应用共词网络可视化、聚类系数、中心度指标、结构洞识别等一系列统计分析方法与指标，对一手证据展开解析的基本程序。本研究所展开的方法，有助于为完成数字人文“一手证据”的循证范式提供参考。

关键词：循证数字人文；一手证据；共词分析；《鲍氏国策》

中图分类号：G250.7　　文献标识码：A　　文章编号：1002-1248（2022）11-0014-12

引用本文：魏志鹏，赵悦言，杨克虎，等. 数字人文“一手证据”循证范式研究：基于《鲍氏国策》的共词分析[J]. 农业图书情报学报，2022，34(11)：14-25.

1　背　景

证据的获取是影响循证数字人文研究的最关键因素之一。按照证据来源于原始文献还是“二手”的研究文献，可以将循证数字人文研究的证据分为“一手证据”和“二手证据”。本文中，所谓的“一手证据”是指来自于原始文献的证据及其关联要素。

西汉刘向在校理皇家书库时，对所见的 6 种记载战国纵横家说辞的作品，包括《国策》《国事》《事语》《短长》《长语》《修书》，进行了编撰。因书所记载的多是战国时纵横家为其所辅之国提出的政治和外交策略，因此刘向把这本书名为《战国策》。该书是一部兼具史学和文学特色的传统典籍。该书记载了上

收稿日期：2022–10–30

基金项目：国家社会科学基金重大项目“循证社会科学的理论体系、国际经验与中国路径研究”（19ZDA142）

作者简介：魏志鹏，博士研究生，兰州大学循证社会科学交叉创新实验室、兰州大学基础医学院循证医学中心。赵悦言，硕士研究生，西北师范大学商学院。杨克虎，教授，兰州大学循证社会科学交叉创新实验室、兰州大学基础医学院循证医学中心

* 通信作者：周文杰，教授，西北师范大学商学院，兰州大学循证社会科学交叉创新实验室，兼职教授，研究方向为信息社会问题。Email：wj_lp@sina.com

起前490年智伯灭范氏，下至前221年高渐离以筑击秦始皇，两百多年间策士的游说活动等历史事件。《战国策》比较全面地反映了战国时期各国的政治、军事、外交方面的一些活动情况和社会面貌。《战国策》善于述事明理，文中大量运用寓言、譬喻，语言生动，富于文采，对中国两汉以来史传文政论文的发展都产生过相当的影响[1,2]。

由于成书年代较早，在传承过程中《战国策》多有散佚。南宋文人鲍彪对《战国策》进行了编订并注释，从而形成了《鲍氏国策》。鲍彪，字文虎，龙泉（今属浙江）人（也有一说为缙云壶镇人）。鲍彪潜心注释《战国策》，收集散佚严重的《战国策》，对收集的史料进行了考辩梳理，重新编排了次序和世系，对其进行了点评，四易其稿终完成[3,4]。

本文以西北师范大学图书馆收藏的古籍珍本《鲍氏国策》为研究对象，通过共词分析方法，提取文献中的数字人文元素，并对其结构特征加以有效分析，从而揭示基于古籍文献的“一手证据”展开文献循证的基本路径与主要步骤，进而为研究者基于来自数字化古籍文献的原始证据展开文献循证研究提供方法规范参照。

2 研究设计

2.1 工具与方法

本文使用的分析工具与分析方法主要包括：使用专门的中文古籍文献分词工具Jiayan（甲言）对《鲍氏国策》进行分词与词性标注。使用本团队自主开发的共现矩阵构造代码，构建词语共现矩阵。应用networkX展开度中心性、中介中心度、接近中心度等相关指标的分析。应用本团队优化改良的莱顿算法，展开结构洞及其他关键网络结构特征分析。使用networkX对共词网络进行聚类分析及其他共词网络特征的分析。

2.2 分析步骤

为提取《鲍氏国策》中的证据要素，本研究按照如下步骤展开了研究。

（1）从西北师范大学图书馆的特藏数据库中，下载了《鲍氏国策》全文，以备分析。

（2）应用Jiayan（甲言）库，对《鲍氏国策》（共十卷）中的全部文献进行基于隐马尔科夫算法（Hidden Markov Algorithm，HMM）的分词，并标注词性。

（3）根据同义词相关性原理，识别了语义模糊词，并将其加入停用词表。分词的时候将停用词和模糊词表进行删除处理。

（4）将停用词（包括语义模糊词）及非名词全部删除，仅保留全部名词。

（5）对全部名词进行词频统计，并对具有相同频数的名词，根据词义相关度赋予不同的权重，最终形成进行加权后的名词词频统计结果。

（6）根据词频高低分别制作3个共词矩阵，并应用networkX展开了基于共词网络的各种特性统计分析。

3 研究结果

经过处理，本研究制作了一个高频词矩阵。入选该矩阵中的词的加权词频为60及以上，共有102个关键词。将加权词频为30~60之间的词语识别为中频词，共有92个关键词。将加权词频为20~30的72个关键词确定为低频词。基于这3种频次不同的词语，本研究构建的3个共词矩阵，并展开了统计分析。

3.1 “一手证据”的可视化

在循证数字人文研究中，词频是一个基本的统计指标。很多研究者都希望通过词频分析，帮助读者快速提取文本中的证据要素，把握原始文本的主题。然而，在本团队的前序研究中发现，仅提供词频，并无助于读者提高阅读效率[5]。为此，循证数字人文研究中，需要基于原始文本中的关键词等“一手证据”，展开更进一步，更具有整体性的分析。共词分析通过关键词对共现的情况来展示关键词间的关联强度，从而有助于展示出《鲍氏国策》所记述的内容中，各证据要素之间的关联关系。例如，在共词网络中，从一个

关键词节点发出的连线越多，表示该关键词所表征的证据信息越重要。按照这一原理，本研究对上述 3 个共词矩阵进行了可视化分析。

3.1.1　高频关键词共现网络

本研究中，经过细致的词语筛选，将经过处理后的名词词表作为“一手证据”的表征。为此，高频词的共现网络直观地显示了原始文本中主要“一手证据”及其与其他证据要素之间的关联。图 1 是对《鲍氏国策》中，最高频的 102 个名词所构造的共词网络。

由图 1 可见，“楚王”“罪”“问”“氏”等节点周围连线较为密集，处于网络核心位置；而“国”“女”“分”等节点连线相较于其他节点较为稀疏，处于边缘位置。这些信息，能够为数字人文研究者提供一些重要的循证线索。也就是说，循证数字人文研究中，可以根据这些最高频次的名词之间的关联特征，分析“一手证据”的主要元素。

3.1.2　中频关键词共现网络

上文对《鲍氏国策》中最高频次的 102 个词进行了共词网络分析，从而提供了一个数字人文研究的循证渠道。进而，本研究对词频低于上述 102 个词，但仍然相对比较高频次出现的 92 个关键词构建了如下图所示的共词网络。

由图 2 可见，“百姓”“大夫”“群臣”等节点周围连线较为密集，处于网络核心位置；而“攻楚”“横”等节点连线相较于其他节点较为稀疏，处于边缘位置。与图 1 不同，图 2 中的名词出现次数并非很高，但其网络仍然有疏有密。根据图中词语之间的关联关系，数字人文研究者就可以进一步获取更多关于“一手证据”的信息。

3.1.3　低频关键词共现网络

高频与中低的词语为数字人文研究者提供了一定的证据信息。在此基础上，本研究针对词频相对较低的名词进行了共词网络分析。

由图 3 可得，相较于高、中频关键词网络，低频关键词无论是在节点个数和边的数量上都较为稀疏，有更多的边缘节点。通过对上述 3 类网络中不同词语及词语类关系的比较分析，数字人文研究者就可以得到很多基本的“一手证据”信息，从而为更加全面、细致的循证研究奠定基础。

3.2　“一手证据”的网络统计特性

基于共词矩阵，循证数字人文研究者可以使用网络节点数、网络关系数、最短路径、全局效率和聚类系数等统计指标，对“一手证据”的统计特性进行了进一步的深入分析。表 1 中，展示了本研究基于《鲍氏国策》所构建的 3 类频数不同的名词网络在上述 5

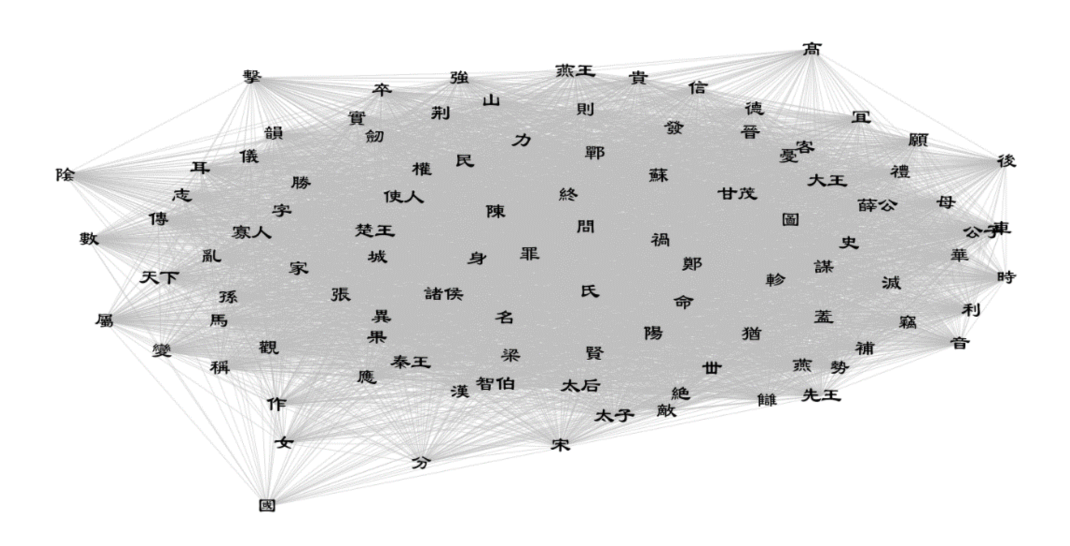


图 1　词频大于 100 的名词共词网络

Fig.1 Noun co-word network with word frequency greater than 100

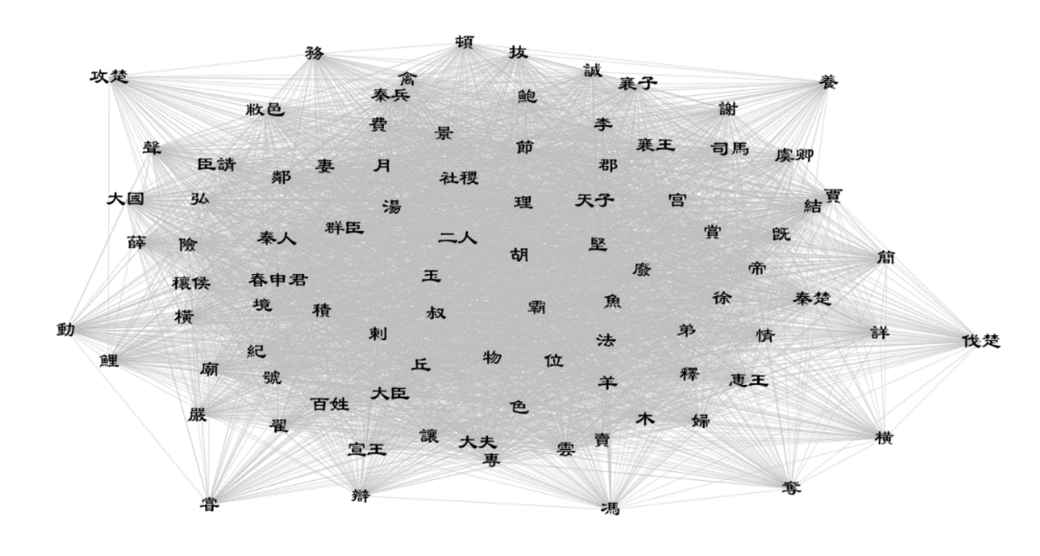


图 2　词频在 30~60 之间的名词共词网络

Fig.2 Noun co-word network with word frequency from 30 to 60

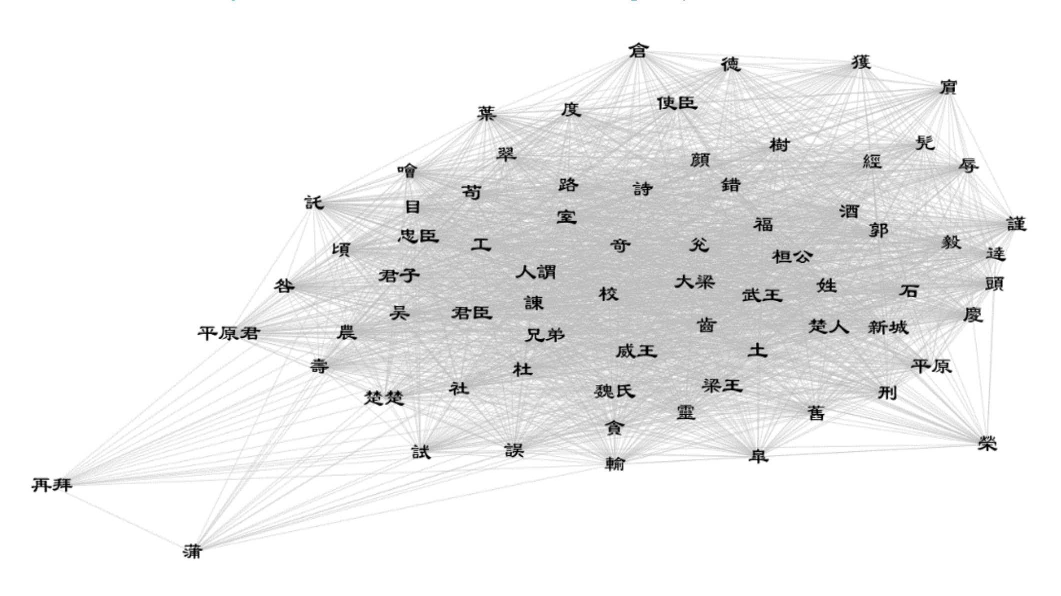


图 3　词频在 20~30 之间的名词共词网络

Fig.3 Noun co-word network with word frequency from 20 to 30

表 1　高、中、低 3 个共词网络的五种统计指标

Table 1 Five statistical indicators of high-, medium- and low-frequency co-word networks

比较对象	网络节点数/个	网络关系数/个	最短路径	聚类系数	全局效率
高词频共现网络	102	19 229	1.061 93	0.954 2	0.969 04
中词频共现网络	92	12 389	1.083 61	0.936 1	0.958 19
低词频共现网络	72	5 599	1.138 11	0.910 7	0.930 95

个指标上的统计结果。

3.2.1　网络密度及聚类系数分析

由表 1 可看到，随着高、中、低词频共现网络的网络节点数依次递减，关键词间的共现关系数也处于递减状态，说明两者之间存在着关联关系。同时，3 个网络的聚类系数都高于 0.9，说明 3 个网络都具有很好

的凝聚力，其中高词频共现网络的聚类系数为 0.954 2 为最大，证明高词频共现网络凝聚力最高。这 3 个网络均具有略大于 1 的平均最短距离和 0.9 以上的聚类系数、说明关键词间具有一定的关系、存在网络小世界效应（小世界网络具有平均最短路径长度小、聚类系数大的特点）；从一定程度上也反映出高、中、低词频关键词间均具有良好的关联性、网络连通性也较好。

3.2.2 最短路径及全局效率分析

最短路径是连接任何两个关键词之间的最短途径的平均长度，平均距离短表示关键词间信息传播快，影响大；最短路径越短表示关键词间信息传播快，影响大。全局平均效率是所有结点对的平均效率，建立在“距离”基础上用来测量网络的整体凝聚力[6]。由表 1 可以看出，与中词频及低词频相比，高频词网络中关键词传播最快，效率最高，这表明高词频网络具有最佳的整体凝聚力和关联程度。

3.2.3 中心性分析与比较

中心性用来反映各个关键词在共词网络中所在中心的程度，测度方法主要分为点度中心度、接近中心度、特征向量中心度和中介中心度 4 种。

点度中心度（Degree Centrality）是刻画中心性最直接最直观的测度指标，是该关键词在共现网络中与其他关键词存在联系的节点数除以 n-1（其中 n-1 就是归一化使用的常量），点度中心度越大，说明该关键词与其他关键词产生的联系越多，即该关键词在共现网络中的地位就越高。

中介中心度（Betweenness Centrality）指的是一个关键词担任其它两个关键词之间最短路的桥梁的次数，一个关键词充当“中介”的次数越高，它的中介中心度就越大。中介中心性用来衡量哪些关键词具有良好的沟通和信息传递的桥梁作用。

特征向量中心度（Eigenvector Centrality）认为，一个关键词的重要性既取决于其邻居词的数量（即该节点的度），也取决于其邻居词节点的重要性。特征向量中心度算法是一个用来度量节点之间的传递影响和连通性的算法，在特征向量中心度算法中，其认为与具有高得分的节点相连接比与具有低得分的节点相连接所得的贡献更大[7]。

按照接近中心度（Closeness Centrality）的原理，一个关键词的接近中心度较高，说明该关键词到网络中其他各关键词的距离总体来说较近。如果节点到图中其他节点的最短距离都很小，那么它的接近中心性就很高。相比中介中心性，接近中心性更接近几何上的中心位置。

根据 LEE 的研究[8]，共词网络中关键词的中心性可以用来衡量一个关键词在网络中的重要性。基于上文所述的 4 个关键性指标，本研究应用 Python 软件中的 networkX，计算了《鲍氏国策》中主要关键词在 4 个中心度上的值。为节约篇幅，本研究仅列出上述 4 项中心度指标排在前 10 的关键词（部分关键词指标取值相同，则只取一个值且在关键词合并表示）。具体结果如表 2 所示。

（1）高词频网络中心度分析。如上文所述，点度中心性用来衡量在网络中哪些关键词最重要。由表 2 可见，在高频词中，“勢、氏、命、力、民、楚王、身、利、字、志、燕、諸侯、天下、史、家”等关键词处在网络的最核心位置，点度中心度为 1，说明上述关键词与其他关键词均有联系，在网络中处于最重要的位置。表 2 中其余高频关键词的点度中心度均大于等于 0.970 3，说明表中的这些高频关键词也处在网络中较重要位置。中介中心性用来衡量哪些关键词具有桥梁作用，具有传递和沟通作用。由表 2 高频词可见，“勢、氏、命、力、民、楚王、身、利、字、志、燕、諸侯、天下、史、家”和“鶩、滅”等关键词中介中心性为表中最高，具有最强影响力。同时需要注意的是，上述这些词的中介中心性最大值为 0.001。由此可见，高词频网络中具有桥梁作用的关键词较少，关键词传递信息和沟通能力较弱。表 2 高频词中特征向量中心性最高的是“勢、氏、命、力、民、楚王、身、利、字、志、燕、諸侯、天下、史、家”，说明与这些关键词相连接的邻居节点处于重要位置，但列出的 10 行高频关键词特征向量中心度相差不大，总体偏低且差距不明显。接近中心性用距离来衡量关键词的中心程度。表 2 高频词中，“勢、氏、命、力、民、楚王、

表 2　高、中、低词频网络关键词中心性数值（前 10）

Table 2 Keywords' centrality in high-, medium- and low-frequency co-word network (Top 10)

高词频	Dc	Bc	Ec	Cc
勢、氏、命、力、民、楚王、身、利、字、志、燕、諸侯、天下、史、家	1.000 0	0.001 0	0.104 2	1.000 0
陳、滅	0.990 1	0.001 0	0.103 3	0.990 2
權、謀	0.990 1	0.000 9	0.103 4	0.990 2
公子、果、太子、城、耳、寡人、梁、使人、名、秦王、太后、母、罪、音、客、山、德、陽、荆	0.990 1	0.000 7	0.103 7	0.990 2
禍	0.980 2	0.000 9	0.102 6	0.980 6
禮	0.980 2	0.000 8	0.102 6	0.980 6
智伯、信、卌、應	0.980 2	0.000 6	0.103 0	0.980 6
張	0.970 3	0.000 6	0.102 0	0.971 2
問	0.970 3	0.000 6	0.102 1	0.971 2
車	0.970 3	0.000 5	0.102 1	0.971 2
中词频	Dc	Bc	Ec	Cc
大夫、天子、情、二人、月、帝、境、弟、秦兵、郡、惠王、群臣、丘、徐、霸、胡、色、妻、社稷	1.000 0	0.001 5	0.111 9	1.000 0
百姓、法、羊、玉	0.989 0	0.001 3	0.111 1	0.989 1
宫、刺	0.989 0	0.001 3	0.111 0	0.989 1
物、魚、李	0.989 0	0.001 3	0.111 1	0.989 1
穰侯	0.989 0	0.001 3	0.111 0	0.989 1
婦、位、木	0.978 0	0.001 2	0.110 0	0.978 5
廟、紀	0.978 0	0.001 3	0.110 0	0.978 5
專	0.978 0	0.001 1	0.110 3	0.978 5
誠	0.967 0	0.001 2	0.108 9	0.968 1
釋	0.967 0	0.001 0	0.109 2	0.968 1
低词频	Dc	Bc	Ec	Cc
兄弟、君子、校、人謂、土	1.000 0	0.005 2	0.131 2	1.000 0
君臣	0.985 9	0.004 7	0.130 1	0.986 1
諫	0.985 9	0.005 0	0.129 7	0.986 1
奇、目、路、工、武王、郭、福、慶、室、齒、新城、兗	0.971 8	0.002 4	0.130 1	0.972 6
頃	0.971 8	0.004 5	0.128 4	0.972 6
石、忠臣	0.957 7	0.002 1	0.129 0	0.959 5
桓公	0.957 7	0.002 4	0.128 3	0.959 5
頭	0.957 7	0.002 2	0.128 6	0.959 5
大梁、酒	0.943 7	0.002 2	0.126 8	0.946 7
噲	0.943 7	0.001 9	0.127 3	0.946 7

身、利、字、志、燕、諸侯、天下、史、家”等关键词接近中心性为 1 最大，说明它与其他关键词之间的距离最小可以和其他所有关键词直接联系且联系较为紧密。

（2）中词频网络中心度分析。由表 2 中频词结果可见，“大夫、天子、情、二人、月、帝、境、弟、秦兵、郡、惠王、群臣、丘、徐、霸、胡、色、妻、社稷”等词语的点度中心性最高，数值为 1，其余关键词的点度中心度均高于 0.967 0。这表明，表 2 中的中频关键词均处在网络中较重要位置。其中“大夫、天子、情、二人、月、帝、境、弟、秦兵、郡、惠王、群臣、丘、徐、霸、胡、色、妻、社稷”等关键词中

介中心性最高，具有最强传递和沟通能力。由于这些词的中介中心性最大值仅为 0.001 5，由此可见，中词频网络中几乎没有具有桥梁作用的关键词。表 2 中中频词特征向量中心性和接近中心性最高的词是“大夫、天子、情、二人、月、帝、境、弟、秦兵、郡、惠王、群臣、丘、徐、霸、胡、色、妻、社稷”，说明与这些关键词相连接的邻居节点处于重要位置，且这些词语与其他所有关键词之间存在直接联系。

（3）低词频网络中心度分析。由表 2 低频词结果可见，“兄弟、君子、校、人謂、土”这 5 个关键词在 4 个中心性指标上均为最高，这说明这 5 个低频关键词不仅处于网络中最重要的位置，也拥有着最强的传递及沟通能力，并且与其相连接的邻居节点也处于重要位置。整体而言，这 5 个词与其他词之间具有紧密联系。

综上所述，基于 4 种中心性数据比较分析，数字人文的研究者可以充分挖掘原始文本中的证据信息，并结合具体的情境知识，对这些证据做出进一步深入分析。

3.2.4 结构洞分析

结构洞是指网络中的某些个体与其他个体有直接联系，但也与另一些个体不发生直接联系或关系间断，从而在整体上看好像网络结构中出现了洞穴的现象[9]。结构洞主要用来衡量网络连接中的空洞位置的大小。目前学界内对结构洞的测量主要通过有效规模（Effect Size）、效率（Efficiency）、限制度（Constraint）3 种方式。本研究采用限制度和有效规模这两对指标对结构洞进行衡量。

限制度指网络中利用结构洞的受限程度。具体而言，是指自我节点与其他节点相连所受到的约束程度，与节点可获取权利呈反向关系，约束越高，可获取的资源和信息越少。限制度表征了个体网络的闭合性，即网络中自我节点与其他节点直接或间接的紧密程度，限制度越低，与相连的其他节点所覆盖的网络越开放，网络越富含结构洞，越富有信息利益和控制利益。

有效规模用来测量节点在网络中的非冗余因素。具体而言，某个节点的有效规模等于该节点的个体网络规模减去网络的冗余度。节点的有效规模实际上就是其邻域中所包含的节点的数量，冗余度则等于该节点所在的个体网络成员中其他节点的平均度数。所以，有效规模等于个体网络规模减去此节点该个全网络成员的平均度[10]。有效规模可以测算某一关键词的整体影响力。有效规模越大，表明这些关键词在整个网络中因与其他关键词有诸多连线，从而具有重要控制力和影响力。

（1）高频关键词。从表 3 中可以看出，在高频关键词网络中有效规模最高的关键词为“氏、民、身、燕、勢、命、力、楚王、利、字、志、諸侯、天下、史、家”，这些关键词的有效规模值为 7.316 8。由此表明这些词在整个网络中与其他关键词连线很多，因此具有重要控制力和影响力。有效规模越大，说明节点在网络中地位越重要，反之亦然。按此标准判断，“智伯、世、應”等词在网络中地位重要性较小。

在共现网络中，有些关键词只与另外一个关键词相连，从而受与其相联系的其他词的严格约束。这导致了这些词语对其他词语的依赖性大，跨跃结构洞能力极弱。由表 3 可见，“氏、民、身、燕、勢、命、力、楚王、利、字、志、諸侯、天下、史、家、滅、鶩”等词的约束度为 0.039 3，表明在高词频共现网络中，这些关键词与较多关键词相连，其受与其相联系词的约束性和依赖性最小，其跨跃结构洞能力极强。

（2）中频关键词。与上述分析同理，在中频关键词网络中有效规模最高，约束度最小的关键词包括“大夫、天子、情、二人、月、帝、境、弟、秦兵、郡、惠王、群臣、丘、徐、霸、胡、色、妻、社稷”，由此表明，这些词在网络中地位最重要且受到的约束最小，可获得更多的资源与信息。

（3）低频关键词。在低频关键词网络中有效规模最高，约束度最小的关键词为“兄弟、君子、校、人謂、土”。而在表 3 中有效规模最低，约束度最大的关键词为“大梁、酒”。由此可见，“大梁、酒”和“兄弟、君子、校、人謂、土”等词语相比，在网络地位重要性较小，受到的约束性和依赖性较大，其跨跃结构洞能力较弱。

表 3　高、中、低词频网络关键词限制度和有效规模（前 10）

Table 3 Constraint and effective scale of keywords in high-, medium- and low-frequency co-word networks (Top 10)

高词频	约束度	有效规模
氏、民、身、燕、勢、命、力、楚王、利、字、志、諸侯、天下、史、家	0.039 3	7.316 8
滅	0.039 3	7.060 0
陳	0.039 3	6.980 0
權、謀	0.039 4	6.760 0
禍	0.039 5	6.252 5
公子、果、太子、城、耳、寡人、梁、使人、名、秦王、太后、母、罪、音、客、山、德、陽、荆	0.039 5	6.160 0
禮	0.039 5	6.151 5
孫	0.039 5	6.051 5
先王	0.039 6	5.783 5
智伯、世、應	0.039 6	5.606 1
中词频	约束度	有效规模
大夫、天子、情、二人、月、帝、境、弟、秦兵、郡、惠王、群臣、丘、徐、霸、胡、色、妻、社稷	0.043 6	8.692 3
宫、刺、穰侯	0.043 7	8.022 2
百姓、法、羊、玉、物、魚、李	0.043 8	7.888 9
廟、紀	0.043 8	7.651 7
婦	0.043 8	7.561 8
木、位	0.043 8	7.516 9
理	0.043 9	7.344 8
誠	0.043 9	7.227 3
專	0.043 9	7.157 3
湯	0.044 0	6.931 8
低词频	约束度	有效规模
兄弟、君子、校、人謂、土	0.055 8	10.943 7
諫	0.056 0	10.485 7
君臣	0.056 1	10.114 3
頃	0.056 2	9.753 6
苟	0.056 7	8.656 3
奇、目、路、工、武王、郭、福、慶、室、齒、新城、兌	0.056 8	8.217 4
桓公	0.056 8	8.029 4
梁王	0.057 0	8.060 6
頭	0.057 0	7.735 3
大梁、酒	0.057 0	7.537 3

3.3　“一手证据”的聚类分析

聚类系数和节点紧密性活力是两个衡量共词网络节点间关系密切程度的指标。聚类系数（Clustering Coefficient，CC）衡量了特定节点的单跳邻居之间是否相关关联。也就是说，如果节点 a 和节点 b 连接，并且 b 和 c 连接，那么 a 和 c 也有连接的可能性。一个节点的局部聚类系数体现的是其邻节点也相互连通的可能性。聚类系数越大说明存在较多的紧密联系团体，简洁性低。节点的紧密性活力（Closeness Vitality，CV）则统计了排除特定节点后，所有其他节点对之间距离之和。

由表 4 可见，3 个共词网络中，高频词的聚类系数最大，代表这些节点的网络粘性高。相比而言，中频词和低频词差距不大。整体而言，3 个网络都属于凝聚力强，联系紧密的网络。根据表 4 节点的聚类系数和紧密型活力分析，可以看出，高频关键词“國、鬻、分、女、高、擊、憂、數”与中词频关键词“攻楚、富、伐楚、奪、横、嚴”和低词频关键词“再拜、蒲、桊、玙、倉、賔、德、平原君、阜”同其他关键词联系紧密，排除它们时网络会发生很大的改变。

4 讨　论

随着自然语言处理技术的发展，数字人文领域的很多研究者都致力于发展新的统计技术，以改进数字人文研究的效率。对于循证数字人文研究来说，主题建模方法是当前数字人文研究中一种新的计算方法。王小红等[11]以主题建模在人文知识研究和学习中的应用为例，剖析了主题建模对人文知识的计算分析所引发的主题客观性呈现、解读语义、新的阅读方式等方法论、认识论问题。

除主题建模外，也有研究者[12]基于中国历代人物资料库（CBDB）数据库，构建历史人物关系网络，将历史人物表示成具有语义的低维实向量。这些研究基于数字人文理念，对人物相关度计算和人物关系挖掘等人文计算任务展开了实证研究，为数字人文中基于“一手证据”而展开循证研究提供了可行的方案。此外，张云中等[13]针对唐三彩数字文化资源展示设计了语义描述模型与元数据框架。刘浏等[14]从数字人文视角入手，解析了古汉语实体歧义问题。朱锁玲等[15]就数字人文在中国农史研究中的实践展开了研究。魏晓萍[16]则针对数字人文背景下数字化古籍的深度开发利用进行了探析。

另有研究者[17]通过抽取古籍文献中蕴含的物产名与别名的关联关系为数据对象，借助社会网络分析技术，通过线值、点度、个人中心网络、连通子网络等维度，可视化地展示物产名与别名之间的网络关系。进而从不同的视角进行知识关联分析，探讨社会网络分析技术在方志类古籍知识挖掘中的应用。研究表明，社会网络分析方法在古籍的知识挖掘方面有良好的应用效果。此外，王丽丽等[18]对数字人文视角下古籍知识关联相关研究进行了梳理，认为借助于数据分析技术、机器学习、可视化技术等可实现古籍知识关联。并提出古籍知识关联起点是文献组织，基础是古籍数据，本质是知识组织。程结晶等[19]以《汉书·艺文志》中西汉经学家群体为研究对象，在数据资源的数据分类、实体属性阐释、词表构建以及本体模型的确定的基础上，搭建相关知识关联的组织框架，并对知识组织框架中的源数据层、数据转换层、数据关联层、知

表 4　高、中、低词频网络关键词聚类系数和节点紧密性系数（前 10）

Table 4 Clustering coefficient and closeness vitality of high-, medium- and low-frequency co-word networks (Top 10)

高频词				中频词				低频词			
CC		CV		CC		CV		CC		CV	
女	1.000 0	國	162	攻楚	1.000 0	攻楚	131	獲	1.000 0	再拜	123
分	1.000 0	陰	133	伐楚	1.000 0	嘗	128	倉	1.000 0	蒲	123
擊	0.993 7	分	133	嘗	1.000 0	養	125	賔	1.000 0	桊	105
高	0.985 9	女	133	養	0.993 7	伐楚	125	再拜	1.000 0	獲	98
陰	0.985 5	高	132	奪	0.986 3	奪	121	蒲	1.000 0	倉	98
國	0.982 1	擊	128	横	0.984 7	横	120	桊	0.986 5	賔	98
燕王	0.979 2	作	121	抜	0.981 9	動	114	德	0.978 7	德	94
憂	0.978 7	憂	121	嚴	0.976 7	馮	112	平原君	0.971 4	平原君	92
甘茂	0.978 0	數	117	務	0.971 8	嚴	111	楚楚	0.963 0	試	91
數	0.974 2	變	115	辯	0.970 3	鯉	111	阜	0.960 2	阜	88

识应用层进行阐释，试图为古籍中人物史料的语义化组织提供可参考的研究渠道。马创新等[20]提出使用结构化的知识表示方法，组织经典古籍和注疏文献中的知识，实现知识的自动重组和聚类，分析注疏文献中存在的问题。

综上所述，数字人文的研究者已针对原始古籍文献中所蕴含的丰富证据元素展开了大量研究。通过系统性文献调查发现，基于“一手证据”的数字人文循证范式正在形成。同时，在文献调查中也可以看出，循证数字人文研究者对于“一手证据”的挖掘相关研究仍然比较零散，对于构建具有中国特色的数字人文循证范式自主体系而言，加强基础研究，将目前较为散乱的指标整合为系统化的方法体系已刻不容缓。基于此，本研究以《鲍氏国策》为研究对象，通过共词分析的方法，为数字人文循证研究范式的形成做出了有益的探索。

5 结 语

面对丰富多样的中华典籍文献，构建具有中国特色数字人文循证范式自主知识体系具有很高的可行性和理论与实践价值。本文以《鲍氏国策》为例，基于自然语言处理技术，以共词分析方法为突破口，比较全面地展开了数字人文研究者如何基于来自原始文献的“一手证据”展开系统化的研究，并提供了相应的统计分析指标。面对方兴未艾的数字人文研究，构建具有中国特色、中国风格、中国气派的循证数字人文学科、学术和话语体系，可谓恰逢其时。基于本研究所发展的循证数字人文“一手证据”分析、挖掘的思路和方法，在后续研究中，我们将把更多的优秀传统经典纳入其中，发展出更加系统化的分析方法，从而使循证数字人文研究彰显出最大的研究效益。

在本团队展开的前序研究中，已展示了循证社会科学研究中，科学展开一手文献证据检索的原理与方法[21,22]。本研究进一步比较全面地展示了如何通过自然语言处理等方法，从原始的典籍文献中获取丰富的“一手证据”，以便展开循证数字人文研究。然而，面对海量的典籍、档案、简牍等宝贵文化遗产，如何应用先进的技术工具，进一步展开证据的挖掘与关联分析，仍然是一个亟待开拓的崭新领域。为此，本研究的开展，为开展更大规模的“一手证据”循证数字人文研究提供了一个可资借鉴的案例。期待在今后的数字人文研究中，更多具有中国特色的数字人文研究不断涌现，从而为早日建成中国自主的循证数字人文知识体系奠定基础。

参考文献：

[1] 徐中舒. 论《战国策》的编写及有关苏秦诸问题[J]. 历史研究, 1964(1): 133-150.
XU Z S. On the compilation of warring states policy and some problems related to Su Qin[J]. Historical research, 1964(1): 133-150.
[2] 頡刚. 戰国策之古本与今本[J]. 历史研究, 1957(9): 32.
JIE G. Ancient and present versions of warring states policy[J]. Historical research, 1957(9): 32.
[3] 吴怀东, 徐昕. 宋代杜诗注家鲍彪考[J]. 杜甫研究学刊, 2014(1): 80-86.
WU H D, XU X. Textual research on Bao Biao, an annotator of Du Fu's poems[J]. Journal of Dufu studies, 2014(1): 80-8
[4] 霍旭东. 宋元时期整理《战国策》的巨大成就——兼对鲍彪整理《战国策》再评价[J]. 烟台大学学报(哲学社会科学版), 1989(2): 57-64.
HUO X D. Great achievements in sorting out the warring states policy in Song and Yuan dynasties - Re-evaluation of Bao Biao's sorting out the warring states policy[J]. Journal of Yantai university (philosophy and social science edition), 1989(2): 57-64.
[5] 周文杰. 数字信息分析的辅助策略实验研究: 基于高频词及其可视化呈现[J]. 图书情报工作, 2011, 55(24): 48-51.
ZHOU W J. Auxiliary-strategies of users' digital information analysis: Based on providing of high-frequency-word lists and their visulization[J]. Library and information service, 2011, 55(24): 48-51.
[6] LATORA V, MARCHIORI M. Efficient behavior of small-world networks[J]. Physical review letters, 2001, 87(19): 198701.
[7] RACA V, CICO B. Social network analysis, methods and measurements calculations[C]. 2013 2nd mediterranean conference on em-

bedded computing (MECO), 2013: 251-254.

[8] LEE W H. How to identify emerging research fields using scientometrics: An example in the field of information security[J]. Scientometrics, 2008, 57(3): 357-377.

[9] BURT R S. The social origins of good ideas[EB/OL]. [2023-01-01]. http://www.analytictech.com/mb709/readings/burt_SOGI.pdf.

[10] 王国明, 李夏苗, 胡正东, 等. 长株潭城市群交通网络结构洞分析[J]. 计算机工程与应用, 2012, 48(15): 1-6.

WANG G M, LI X M, HU Z D, et al. Study on traffic networks of urban agglomeration of Chang-Zhu-Tan based on structural holes theory[J]. Computer engineering and applications, 2012, 48(15): 1-6.

[11] 王小红, 科林·艾伦, 浦江淮, 等. 人文知识发现的计算机实现——对“汉典古籍”主题建模的实证分析[J]. 自然辩证法通讯, 2018, 40(4) :50-58.

WANG X H, COLIN A, PU J H, et al. To discover humanities knowledge by the computer: An empirical analysis of topic modeling the "Handian" ancient Chinese classics[J]. Journal of dialectics of nature, 2018, 40(4): 50-58.

[12] 潘俊. 面向数字人文的人物分布式语义表示研究——基于 CBDB 数据库和古籍文献[J]. 图书馆杂志, 2020, 39(8): 94-102.

PAN J. Distributed representation learning for the historical figures based on CBDB and ancient books: A digital humanistic perspective[J]. Library journal, 2020, 39(8): 94-102.

[13] 张云中, 焦凤枝, 刘嘉琳. 唐三彩数字文化资源展示的语义描述模型与元数据框架[J]. 图书与情报, 2021(3): 87-96.

ZHANG Y Z, JIAO F Z, LIU J L. The semantic description model for the display of tang tri-color digital cultural resources and metadata framework[J]. Library & information, 2021(3): 87-96.

[14] 刘浏, 王东波, 黄水清, 等. 数字人文视野下的古汉语实体歧义研究[J]. 图书与情报, 2020(5): 115-124.

LIU L, WANG D B, HUANG S Q, et al. Research on ancient Chinese entity ambiguity in digital humanities[J]. Library & information, 2020 (5): 115-124.

[15] 朱锁玲, 包平. 数字人文在中国农史研究中的实践与思考——以中华农业文明研究院数字人文项目为例[J]. 农业图书情报学报, 2021, 33(8): 79-87.

ZHU S L, BAO P. Practice and thoughts on digital humanities in the research of Chinese agricultural history: Taking the digital humanities project of the Chinese academy of agricultural civilization as an example[J]. Journal of library and information science in agriculture, 2021, 33(8): 79-87.

[16] 魏晓萍. 数字人文背景下数字化古籍的深度开发利用[J]. 农业图书情报学刊, 2018, 30(9): 106-110.

WEI X P. Deep development and utilization of digital ancient books under the background of digital humanity[J]. Journal of library and information science in agriculture, 2018, 30(9): 106-110.

[17] 李娜, 包平. 方志类古籍中物产名与别名关系的可视化——基于社会网络分析技术视角[J]. 图书馆论坛, 2017, 37(12): 108-114.

LI N, BAO P. Visual exploration of the relationship between produce names and their alias in ancient local chronicles[J]. Library tribune, 2017, 37(12): 108-114.

[18] 王丽丽, 张宁. 数字人文视角下的古籍知识关联探析[J]. 农业图书情报学报, 2022, 34(9): 51-59.

WANG L L, ZHANG N. An analysis of knowledge correlation of ancient books from the perspective of digital humanity[J]. Journal of library and information science in agriculture, 2022, 34(9): 51-59.

[19] 程结晶, 王璞钰. 古籍中人物史料的关联组织研究——以《汉书·艺文志》中西汉经学家群体为例[J/OL]. 图书馆论坛:1-12[2023-01-01]. http://kns.cnki.net/kcms/detail/44.1306.G2.20211119.1535.010.html.

CHENG J J, WANG P Y. Research on the association organization of historical materials of characters in ancient books - Taking the Western Han Confucian classics group in Han Shu Yi Wen Zhi as an Example[J/OL]. Library tribune: 1-12[2023-01-01]. http://kns.cnki.net/kcms/detail/44.1306.G2.20211119.1535.010.html.

[20] 马创新, 陈小荷, 曲维光. 经典古籍注疏文献的知识网络研究与设计[J]. 图书情报工作, 2013, 57(9): 124-128.

MA C X, CHEN X H, QU W G. Research and design of knowledge network for annotated documents of classical ancient books [J]. Library and information service, 2013, 57(9): 124-128.

[21] 周文杰, 赵悦言, 魏志鹏, 等. 循证视角下文献证据检索的科学性评价: 缘起、指标与趋势[J]. 图书与情报, 2021(6): 31-36.

ZHOU W J, ZHAO Y Y, WEI Z P, et al. Scientific evaluation of literature evidence retrieval quality from the perspective of evidence-

based research: Initiation, index and trend[J]. Library and information, 2021(6): 31-36.

[22] 卢洁妤, 魏志鹏, 周文杰, 等. 文献证据检索的信度研究: 基于循证视角[J]. 图书与情报, 2021(6): 60-68.
LU J Y, WEI Z P, ZHOU W J, et al. Research on reliability of documentary evidence retrieval: Based on evidence-based perspective[J]. Library and information, 2021(6): 60-68.

Evidence-based Digital Humanity Paradigm of First-hand Evidence: A Co Word Analysis Based on Bao's Zhan Guo Ce

WeiZhipeng[1,2], Zhao Yueyan[3], Yang Kehu[1,2], Zhou Wenjie[1,3*]

(1. Cross-innovation Laboratory of Evidence-based Social Science of Lanzhou University, Lanzhou 730030;

2. Evidence-based Medical Center of School of Basic Medical Sciences of Lanzhou University, Lanzhou 730030;

3. Business School of Norwest Normal University, Lanzhou 730070)

Abstract: [Purpose/Significance] Evidence acquisition is one of the most critical factors affecting evidence-based digital humanities research. The first-hand evidence contained in the ancient literature works is an important way to carry out digital humanities research, and thus, the purpose of this research is to shed light on the evidence-based digital humanities research process based on the empirical analysis of Bao's Zhan Guo Ce, which is one of the most influential books in Chinese history. [Method/Process] In the face of rich and diverse Chinese ancient literature works, it is of theoretical and practical value to build an independent knowledge system with Chinese characteristics based on the evidence-based paradigm of digital humanities. For this reason, the present research used the natural language processing (NLP) method to analyze Bao's Zhan Guo Ce in Jiayan Library, which is tailored for the NLP analysis of Chinese ancient literature works. By using co-word analysis, this research comprehensively discusses how digital humanities researchers carry out systematic research based on first-hand evidence from ancient literature via word frequency analysis, visualization of co-words, cluster analysis, centrality degree analysis, etc. Social network analysis (SNA), NetworkX algorithm and co-word visualization procedure are applied to give us insight into how to extract the first-hand evidence from ancient literature works. [Results/Conclusions] The key results include a procedure on how to extract first-hand evidence from ancient literature works like Bao's Zhan Guo Ce, in digital humanities research via Python. Specifically, the procedure includes basic word frequency indicators, a tool of removal of stop words, process of recognition and removal of ambiguous words. Furthermore, this study also takes Bao's Zhan Guo Ce as an example to show the basic procedure of analyzing first-hand evidence in digital humanities research by using a series of statistical analysis methods and indicators such as co-word network visualization, clustering coefficient, centrality degree, and structural hole recognition. The procedures, tools and methods demonstrated in this study are expected to provide reference for completing the evidence-based digital humanity research paradigm of first-hand evidence. Thus, the procedures, tools, statistical indicators and algorithm demonstrated in this research are expected to provide a foundation for building an independent knowledge system of evidence-based digital humanities with Chinese characteristics.

Keywords: evidence-based digital humanity; first-hand evidence; co-word analysis; Bao's Zhan